An Empirical Approach to Credit Risk Modelling

Arunangshu Podder
Great Learning, Bengaluru, India
arunangshu.podder@gmail.com

Nikhil Gupta
Great Learning, Bengaluru, India
nikhil.gupta001009@gmail.com

Madhvi Gaur
Great Learning, Bengaluru, India
madhvigaur@gmail.com

Garima Bajpai
Great Learning, Bengaluru, India
garimabajpai55@gmail.com

***Abstract* — Credit Risk is defined as the risk that is present when a lending institute lends money to an organization or individual. Although, this risk proves to be very profitable if the lent money is paid back with full interest, it can lead to financial disaster in the event of a default. Hence, Credit Risk Modelling becomes a matter of utmost importance for the lending institute in order to mitigate such losses.**

In this paper, we have designed algorithms that can go through the different a borrowers personal information, credit history, details of the loan, etc. and predict whether the borrower will default on the existing loan and in the event of a default, how much of the loan amount is at risk for the lender.

Keywords — Loan, Default Prediction, Loss Given Default, Machine Learning, Regression, Classification

I. INTRODUCTION

The main objective of this project is to identify the factors that influence a loan getting repaid successfully or closed on account of default, based on those factors, build a model that can predict whether a borrower will eventually default on his loan or not, and build separate model that will, in the event of a default, predict the net impact of the default on the lender.

Through this project, we have aimed to help individual lenders or lending institutions minimize the risk of losing money due to bad loans.

II. LITERATURE SURVEY

Loans are very important for any financial institution and are one of the primary sources of income. When a loan goes bad, it can be very fatal for the institution's financial stability. Authors in [1] have shown how bad loans (also known as NPA or Non-performing assets) negatively affects the profitability of banks. All these have given rise to the need of Credit Risk Modelling which is the process of using data models to find out:

a. the probability of the borrower defaulting on a loan, and
b. the effect of this default on the financial stability of the lender.

Our current computational capabilities, emerging data analytics techniques, along with the recent need of automating the loan approval process in the lending sector, have greatly improved the century old practice of predicting the risk of default in the lending process. Numerous papers have been published and research has been done to accurately model Credit Risk using various data science techniques.

One of the main factors on which Credit Risk Modelling is dependent is Probability of Default (PD). Authors in [5], [6] and [10] show us a whole range of modelling techniques using data science for determining PD. Research is also done, e.g. [3] and [4], that tell us how different factors affect loan default. However, with so many algorithms now available with us, there is an obvious question regarding which is the most suitable technique for predicting default. [9] Tells us about a few metrics that can be used for comparing the performances of different algorithms for a given dataset.

Another important factor for Credit Risk Modelling is Loss Given Default (LGD). There have been many studies like [13], [14] and [17] with a focused approach on modelling LGD by considering different attributes of the borrower's personal information, credit history and loan information.

Nowadays, we also have different data science tools like Weka, as mentioned in [7] that can be used for the classification process. Also, there are new fields of research that are further improving the modelling process of Credit Risk like Forensic Analytics where electronic data is being used reconstruct or detect financial fraud.

III. DATA

A. *Data Selection*

For this proposed project, we have looked into data published by various Peer to Peer (P2P) lending platforms since it is very difficult to obtain real bank data. P2P platforms, such as Prosper, Lending Club, Kiva, Fynanz provide various online services, thus enabling individuals and small businesses to get hassle-free loans from interested lenders. Amongst them, Lending Club [23] is the largest P2P lending platform in the world.

B. *Data Specifications*

The dataset released by Lending Club contains 2 million+ loan records issued between the years 2007 and 2018. It has 151 variables which include information such as borrower's credit history, personal information (e.g. annual income, years of employment, zip code), loan information (e.g. description, type interest rates, grade), current loan status, etc. It also contains some variables which define how the settlement will be done for defaulted loans. Such information is a knowledge of the future and need to be handled accordingly.

IV. PROCESS OVERVIEW

A. *Architecture*

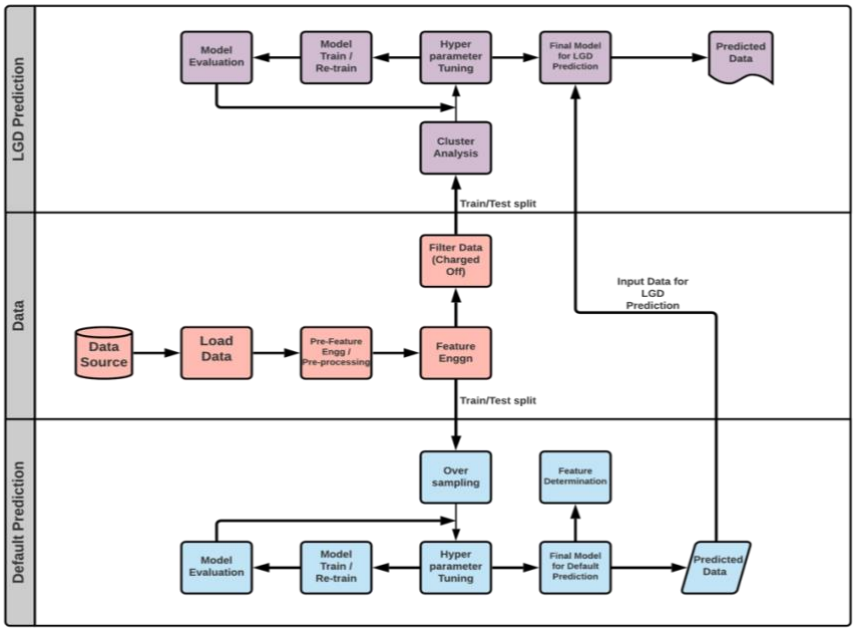

Fig 1

B. *Default Prediction*

For our classification problem, the target variable was "loan_status". As mentioned earlier, the feature had values as:

a. Current: Active Loan.
b. Fully Paid: The full principal with interest rates is paid back.
c. Charged Off: The borrower defaulted on the loan and the loan will never be paid back in full amount.
d. In Grace Period: Payment of installment is delayed by 1 to 15 days.
e. Late (16–30 days): Payment of installment is delayed by 16 to 30 days.
f. Late (31–120 days): Payment of installment is delayed by 31 to 120 days.
g. Default: Payment of installment is delayed by more than 120 days.

We considered a loan with status Fully Paid as "Good Loan" and the one having status Default or Charged Off as "Bad Loan". Lending Club has observed that borrowers who do not pay their due installment for more than 30 days, have ultimately defaulted on their loans. Hence, we considered loans with status Late (31–120 days) as "Bad Loan". This left us with loans Current, In Grace Period and Late (16–30 days), which we considered as "Current Loan" since we cannot exactly say how these loans might end. Thus, we map the different statuses in the dataset as,

a. Good loan: Fully Paid
b. Bad Loan: Charged Off, Late (31–120 days), Default
c. Current Loan: Current, In Grace Period, Late (16–30 days)

For our problem at hand, we considered only the loans with status as Good or Bad to build our model and then used the Current loans to predict which of these active loans are going to default. Hence, we kept our Current loans aside for the time being and focused only on the Good and Bad loans.

After the above-mentioned operation, we found that there was a total of 1369566 records for Good and Bad loans of which approx. 79% are marked as Good Loan and 21.33% are marked as Bad Loan. (Fig 2)

It was quite evident that there was a huge difference in the total number of good and bad loans. So, with this data in hand, whatever model we built would always have been bias towards the Good Loans. Hence, we needed some oversampling technique to remove this imbalance from our dataset.

Two possible over sampling techniques are Random

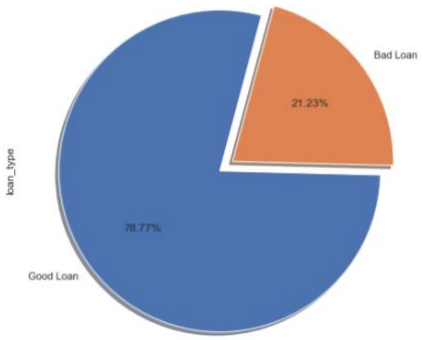

Fig 2

Oversampling and Synthetic Minority Oversampling Technique (SMOTE). SMOTE creates a new sample by observing the closest samples in the feature space. As a result, the performance of SMOTE is too low. Hence, we decided to go with Random Oversampling technique which although just duplicates the minority class, gave us quick results. Also, we compared took 10% of the dataset and compared the performance of both SMOTE and Random Over Sampler and found the performance to be more or less the same.

Once, the dataset was ready, we divided the data keeping 15% for testing and remaining 85% for model training purpose and tried the different classification problems comparing their ROC curves and AUC scores.

C. Loss Given Default Prediction

For building a machine learning model to determine the loss that the lender has to bear in the event of a default, we picked only those samples from the dataset with final status as "Charged Off". We found that to be a total of 2,68,559 records. As discussed in the previous sections, the target variable in this problem was "recoveries".

[13] defines LGD as,

$$LGD = 1 - \frac{total\_payment}{total\_amount}$$

where,

total_payment = total loan amount that has been recovered so far, and

total_amount = total loan amount that should have been recovered.

V. MODEL & RESULTS

A. Default Prediction

Logistic Regression (LR) is defined as a statistical model that uses a logistic function to model a binary classification problem. Here, we are using it to classify between Good Loans which is numerically labelled as 1 and Bad Loans numerically labelled as 0.

The default prediction model has been evaluated on the below metrics.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions\ made}$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

Table I

Evaluation Metric	Value
Accuracy	0.9939
Recall	0.9925
Precision	0.9998
F1	0.9961

The accuracy of our model is 0.9939. This means that during testing about 99.39% of the predictions made by our classification model is correct.

In order to understand how the model performed for each class, we have created the confusion matrix for the model (Fig 3). From this we can see that about approx. 98% of the bad loans were predicted correctly during testing of the model. Similarly, about 99.9% of the good loans are predicted correctly. Thus, we can say that we have built a pretty good classifier, which can predict both the good and bad loans equally well. Further, in order to confirm the performance of our classification model, we look at its ROC score which comes at 0.999 (Fig 4). A value so close to 1 means that we have an excellent classifier in place.

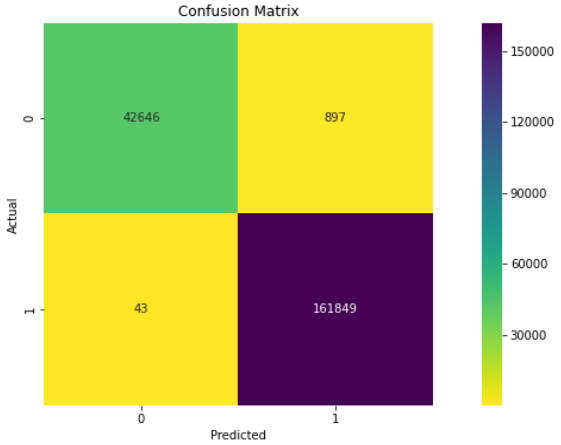

Fig 3

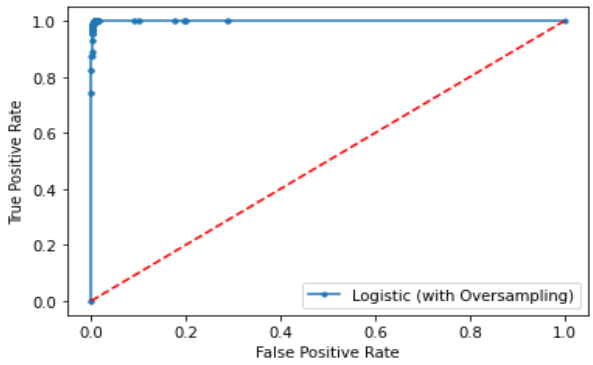

Fig 4

We also looked at the important features in the model. Below (Table II) are the top 5 features deciding whether a loan will end up as a Good Loan or a Bad Loan. A detailed report of the contribution of each feature is given in Appendix A.

Table II

Factors for Good Loans		**Factors for Bad Loans**	
Feature Name	**Feature Importance**	**Feature Name**	**Feature Importance**
total_rec_prncp	22.765821	funded_amnt	-20.220704
last_fico_range_high	1.526253	total_rec_int	-0.991154
term	0.501807	hardship_flag	-0.217384
last_pymnt_d	0.378949	total_rec_late_fee	-0.177522
issue_d	0.220179	total_acc	-0.128451

B. Loss Given Default Prediction

Random Forest Regressor is an ensemble technique that performs regression using multiple decision trees. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees. In this algorithm, we perform random row and column sampling from the dataset forming sample datasets for every model. Below (Table III) illustrated are the results from our experiment.

Table III

Evaluation Metric	**Value**
RMSE	1094.03
R Squared	0.2801
Adjusted R Squared	0.2789

We know that R-Squared determines the percentage of variance in the dependent variable, that the independent variables can explain. Here, our R-Squared value is 0.42 or 42%. That means that only 42% of the variation can be explained by our model.

Since adjusted R-Squared considers only those variables which contribute significantly to our model, we considered the adjusted R-Squared value over the former one. Here, our adjust R-Squared value is almost the same as the R-Squared value.

RMSE is a measure of how spread out these residuals are. In other words, it tells us how concentrated the data is around the line of best fit. Our RMSE score is 977.44. Although compared to other algorithms it is less, still the value is quite high, stating that our model has not done a very good job in predicting the loss that a lender would incur in the event of a default.

One of the primary reasons why our regression model is not able to produce a good result is the skewness that is there in the dependent variable. Because of this skewness (Fig 5), our model is highly lenient towards values closer to 0.

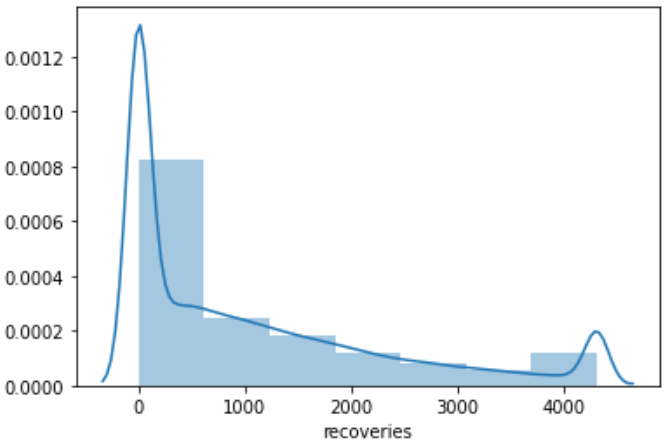

Fig 5

VI. LIMITATIONS & FUTURE WORK

The main limitation in our model is the low accuracy with which it is predicting the loss that a lender will incur in the event of loan default. In real world, sometimes lenders are okay even after knowing the loans that they are funding might default since they expect higher returns. Such low levels of accuracy might be risky for these lenders as in some of these predictions we have seen that the amount at risk was quite high, but the model predicted low amount at risk.

We already analyzed that the main reason for such bad performance by our regression model was due to the skewness that existed in the dependent variable. In order to enhance this, we need to look at ways to remove this skewness.

In [22] we have seen an approach to perform regression for imbalanced dependent variable through an algorithm called SMOGN and this is something we need to look into in order to improve our regression performance.

Also, we need to look into ways to introduce distributed execution through cloud resources while training our model so that we can implement algorithms that were not feasible with our current infrastructure.

REFERENCES

[1] B. Senthil Arasu, P. Sridevi, P. Nageswari, R. Ramya, "A Study on Analysis of Non-Performing Assets and its Impact on Profitability", International Journal of Scientific Research in Multidisciplinary Studies, Volume-5, Issue-6, pp.01-10, June (2019) (d)

[2] Yang Yang, "An empirical study on P2P loan default prediction model", Financial Engineering and Risk Management (2020) 3: 14-22, Clausius Scientific Press, Canada

[3] Michal Polena, Tobias Regner, "Determinants of Borrowers' Default in P2P Lending under Consideration of the Loan Risk Class", Games 2018, 9, 82 (d)

[4] Rajkamal Iyer, Asim Ijaz Khwaja Erzo, F. P. Luttmer, Kelly Shue, "Screening in New Credit Markets: Can Individual Lenders Infer Borrower Creditworthiness in Peer-to-Peer Lending?", SSRN-id1570115 (d)

[5] Don Carmichael, "Modeling Default for Peer-to-Peer Loans", SSRN-id2529240

[6] Alan Zhang, "DEVELOPMENT OF LOGISTIC REGRESSION MODEL TO PREDICT DEFAULT PROBABILITIES OF LOAN ACCOUNTS", International Journal of Information, Business and Management, Vol. 12, No.2, 2020, pg 95-115 (d)

[7] Anand Motwani, Goldi Bajaj, Sushila Mohane, "Predictive Modelling for Credit Risk Detection using Ensemble Method", International Journal of Computer Sciences and Engineering, June 2018, Vol-6, Issue-6 (d)

[8] Zakaria Alomari, Dmitriy Fingerman, "Loan Default Prediction and Identification of Interesting Relations between Attributes of Peer-to-Peer Loan Applications", New Zealand Journal of Computer-Human Interaction ZJCHI 2,2 (2017)

[9] Pedro G. Fonseca and Hugo D. Lopes, "Calibration of Machine Learning Classifiers for Probability of Default Modelling", James Finance (CrowdProcess Inc.), October 24th, 2017 (d)

[10] Rising Odegua, "Predicting Bank Loan Default with Extreme Gradient Boosting" (d)

[11] Jose A. Lopez, Marc R. Saidenberg, "Evaluating Credit Risk Models", Journal of Banking & Finance, Volume 24, Issues 1-2, January 2000, pgs:151-165

[12] Jeremy Turiel, Tomaso Aste, "P2P Loan Acceptance and Default Prediction with Artificial Intelligence", https://www.researchgate.net/publication/334223307

[13] Guangyou Zhou, Yijia Zhang, Sumei Luo, "P2P Network Lending, Loss Given Default and Credit Risks" (d)

[14] Felix Martinson, "Exotic Approaches for Modelling Loss Given Default" (d)

[15] Shunpo Chang, Simon Dae-Oong Kim, Genki Kondo, "Predicting Default Risk of Lending Club Loans"

[16] Carlos Eduardo Canfield Rivera, "Determinants of Default in P2P Lending: The Mexican Case", Independent Journal of Management & Production (IJM&P), v.9, n.1, January - March 2018

[17] Christophe Hurliny, Jérémy Leymariez, Antoine Patinx, "Loss functions for Loss Given Default Model Comparison"

[18] Peter Martey Addo, Dominique Guegan, Bertrand Hassani, "Credit Risk Analysis Using Machine and Deep Learning Models", Risks 2018, 6, 38

[19] Til Schuermann, "What Do We Know About Loss Given Default?", Forthcoming in D. Shimko (ed.), Credit Risk Models and Management 2nd Edition, London, UK: Risk Books, February 2004

[20] Han Sheng Sun and Zi Jin, "Estimating credit risk parameters using ensemble learning methods: an empirical study on loss given default", Journal of Credit Risk 12(3), pgs 43–69

[21] Lin Zhu, Dafeng Qiu, Daji Ergua, Cai Ying, Kuiyi Liu, "A study on predicting loan default based on the random forest algorithm"

[22] Paula Branco, Lu´ıs Torgo, Rita P. Ribeiro, "SMOGN: A Pre-processing Approach for Imbalanced Regression"

[23] https://www.lendingclub.com/

Appendix A

Features deciding Bad Loans	Feature Importance	Feature deciding Good Loans	Feature Importance
funded_amnt	-20.220704	initial_list_status	0.001824
total_rec_int	-0.991154	mths_since_recent_bc	0.003452
hardship_flag	-0.217384	pub_rec_bankruptcies	0.007054
total_rec_late_fee	-0.177522	mths_since_last_delinq	0.008429
total_acc	-0.128451	tax_liens	0.008641
bc_open_to_buy	-0.108019	mo_sin_old_il_acct	0.011208
mo_sin_old_rev_tl_op	-0.104005	num_sats	0.017962
num_bc_tl	-0.073505	mort_acc	0.020196
bc_util	-0.071255	annual_inc	0.021126
last_credit_pull_d	-0.064792	avg_cur_bal	0.029781
fico_range_high	-0.064197	dti_joint	0.031385
open_acc	-0.061272	num_accts_ever_120_pd	0.036386
dti	-0.051507	num_il_tl	0.039755
emp_length	-0.046502	verification_status_joint	0.040833
num_op_rev_tl	-0.042417	acc_open_past_24mths	0.046516
num_tl_90g_dpd_24m	-0.034761	percent_bc_gt_75	0.050051
verification_status	-0.033956	revol_util	0.058122
total_bal_ex_mort	-0.031669	delinq_2yrs	0.060393
num_actv_bc_tl	-0.027890	pct_tl_nvr_dlq	0.063277
home_ownership	-0.027810	total_il_high_credit_limit	0.087392
inq_last_12m	-0.026653	num_rev_accts	0.100289
pub_rec	-0.023027	num_bc_sats	0.123038
annual_inc_joint	-0.021780	sub_grade	0.169488
revol_bal	-0.020074	issue_d	0.220179
purpose	-0.019149	last_pymnt_d	0.378949
mths_since_recent_inq	-0.017456	term	0.501807
total_bc_limit	-0.014192	last_fico_range_high	1.526253
addr_state	-0.013939	total_rec_prncp	22.765821
tot_hi_cred_lim	-0.011188		
collections_12_mths_ex_med	-0.010428		
num_actv_rev_tl	-0.008517		
inq_last_6mths	-0.007061		
total_rev_hi_lim	-0.005645		
chargeoff_within_12_mths	-0.003018		
mo_sin_rcnt_tl	-0.002317		
mo_sin_rcnt_rev_tl_op	-0.001823		
num_tl_30dpd	-0.001146		
delinq_amnt	-0.001125		
tot_coll_amt	-0.000675		

Appendix B

Performance of Different Machine Learning Algorithms (Classification) for Default Prediction:

Model Name	Accuracy	Precision	Recall	F1	AUC Score
Logistic Regression	0.9939	0.9925	0.9998	0.9961	0.999
Support Vector Classifier	0.9922	0.9918	0.9983	0.995	0.998
Random Forest	0.9446	0.9532	0.9356	0.9443	0.986
K Nearest Neighbours	0.9032	0.8943	0.9942	0.9416	0.964
Decision Tree	0.8895	0.923	0.8512	0.8856	0.915

Performance of Different Machine Learning Algorithms (Regression) for Loss Given Default Prediction:

Model Name	RMSE	R^2	Adjusted R^2
Linear Regression	1094.03	0.2801	0.2789
Polynomial Regression	1037.18	0.35	0.3117
Ridge Regression	1051.72	0.2801	0.2789
Lasso Regression	1044.01	0.2799	0.2787
Random Forest Regressor	977.44	0.4254	0.4244